DOCUMENT RESUME

ED 367 678 TM 021 117

AUTHOR Weigle, David C.
TITLE Historical Origins of Contemporary Statistical Testing Practices: How in the World Did Significance Testing Assume Its Current Place in Contemporary Analytic Practice?
PUB DATE Jan 94
NOTE 18p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 27, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Data Analysis; Educational History; Etiology; *Research Methodology; *Scientific Research; *Statistical Significance; *Testing

ABSTRACT

The purposes of the present paper are to address the historical development of statistical significance testing and to briefly examine contemporary practices regarding such testing in the light of these historical origins. Precursors leading to the advent of statistical significance testing are examined as are more recent controversies surrounding the issue. As the etiology of current practice is explored, it will become more apparent whether current practices evolved from deliberative judgment or merely developed from happenstance that has become reified in routine. Examination of the history of analysis suggests that the development of statistical significance testing has indeed involved a degree of deliberative judgment. It may be that the time for significance testing came and went, but there is no doubt that significance testing served as an important catalyst for the growth of science in the 20th century. (Contains 39 references.) (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED 367 678

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
DAVID C. WEIGLE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

# Historical Origins of Contemporary Statistical Testing Practices: How in the World Did Significance Testing Assume Its Current Place in Contemporary Analytic Practice?

David C. Weigle
Texas A & M University
College Station, Texas 77843-4225

Running Head: HISTORY OF SIGNIFICANCE TESTING

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, Texas, January 27, 1994.

TM021117

BEST COPY AVAILABLE

## Abstract

The purposes of the present paper are to address the historical development of statistical significance testing and to briefly examine contemporary practices regarding such testing in the light of these historical origins. Precursors leading to the advent of statistical significance testing are examined as are more recent controversies surrounding the issue. As the etiology of current practice is explored, we may better evaluate whether current practices evolved from deliberative judgment or merely developed as happenstance that has become reified in routine.

The use of statistical significance testing as part of the interpretation of empirical research results has historically generated considerable debate (Carver, 1978; Huberty, 1987; Morrison & Henkel, 1970; Thompson, 1989). A series of articles on the limits of statistical significance testing has even appeared on a seemingly scheduled basis in recent editions of the American Psychologist (Cohen, 1990; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1989). And an entire special issue of the Journal of Experimental Education, published in the Fall of 1993, summarized the numerous criticisms about contemporary reliance on statistical significance testing.

Thompson (1993, p. 285) states that "issues involving statistical significance have probably caused more confusion and controversy than any other aspect of contemporary analytic practice." Yet, despite the continued controversy and debate, "testing for statistical significance has abated very little, if at all" (Carver, 1993, p. 292). The purposes of the present paper are to address the historical development of statistical significance testing and to briefly examine contemporary practices regarding such testing in the light of these historical origins. As the etiology of current practice is explored, we may better evaluate whether current practices evolved from deliberative judgment or merely developed as happenstance that has become reified in routine.

## Precursors of Statistical Significance Testing

Statistical significance testing has been dated to 1900 with Karl Pearson's publication of his chi-square "goodness of fit" test comparing data to a theoretically-expected curve (Gigerenzer *et al.*, 1989). However, the concept of matching hypotheses with data had much earlier beginnings. Perhaps the earliest published report of such a test was that of John Arbuthnot in 1710. Arbuthnot, a physician to

Queen Anne of England, collected data on christenings in London for the period 1629-1710 and found that the number of male births exceeded the number of female births by a small percentage in each of these years. His determination that the probability of this occurrence was extremely small (only $(.5)^{82}$) led him to the conclusion that "it is Art, not Chance, that governs," (Arbuthnot, 1710). In accordance with the deterministic religious philosophy that had held sway from the early Christian era and saw all events as being caused by God rather than by chance, Arbuthnot viewed his findings not only as evidence of God's existence and omnipotence, but, more specifically, as evidence that God had preordained the institution of monogamous marriage.

Further examples of early significance tests include that of John Michell, an English astronomer who, in 1767, published a study asserting that the placement of stars could not be due to chance but were rather due to an unspecified general law--"whether to their mutual gravitation, or to some other law or appointment of the Creator" (Michell, 1767). In 1862 Gustav Kirchoff conducted a similar observational study in which he compared 60 lines of a solar spectrum to 60 lines from a spectrum he produced with iron filings in a Bunsen burner. Kirchoff found that although the probability of his two sets of data actually matching was minute--assuming the null hypothesis was true--they did indeed match.

Pierre Simon Laplace in 1823 tested the hypothesis that the phase of the moon did not influence barometric changes. Laplace (1823) used four separate significance tests based on the comparison of quarterly means from 792 days of data collection. He recognized, perhaps for the first time, the influence of sample size on significance when he reported that the observed mean difference could be confirmed only if it were based on nine times as much data as he had already collected (Stigler, 1986, p. 151).

It is clear that these early precursors of contemporary statistical significance testing did not use modern concepts such as p-values, critical values, etc., and the statistical reasoning involved in their studies was, perhaps, as Gigerenzer *et al.* (1989, p. 89) called it, "window dressing" in relation to the observational evidence. However, a logic consistent with that of contemporary practice laid the foundation for a more extensive development of statistical inference.

Another important precursor of statistical significance testing was the discovery of the normal curve by Abraham De Moivre in 1733 as a byproduct of his method of approximating the sum of a large number of binomial terms. Laplace and Carl Friedrich Gauss further developed applications of the normal distribution in the 1820s. Stigler (1986) notes that the synthesis of the work of Gauss with that of Laplace was "one of the major success stories in the history of science" as it wed the combination of observations through the aggregation of linearized equations of condition with the use of mathematical probability to assess uncertainty and make inferences. Although the applications of this synthesis were widely used in astronomy and geodesy, they remained generally distinct from the social sciences for many years despite their apparent usefulness. With the early efforts at reconciling hypotheses with data and with the advent of the normal distribution, however, the stage was set for the advent of statistical significance testing.

## The Advent of Statistical Significance Testing

Among the earliest statistical significance tests were those that involved the probable error measurement. The term "probable error" was first used by Friedrich Wilhelm Bessel in 1815 (Walker, 1929, p. 186). Probable error refers to "the deviation from a central measure between whose positive and negative values one half the cases may be expected to fall by chance" (Cowles & Davis, 1982, p. 555).

Thus, one probable error unit equals approximately 2/3 of a standard deviation. Bessel used comparisons of probable errors to determine whether a difference was real or due to observational error. Hermann Ebbinghaus, one of the firsts psychologists to apply principles of probability to the measurement of uncertainty, also used probable error to interpret the data he gathered in his important study of memory (Ebbinghaus, 1885). He held that a difference of six times the probable error was fully proven while a difference of twice the probable error was noteworthy. Ebbinghaus also commented that although one could be sure a difference exists when the observed difference is six times the probable error, the observed difference may not be the exact size of the true difference.

Francis Edgeworth, a lawyer and economist who was self-educated in mathematics and statistics, developed a test of significance in which he compared the difference of the means with the "modulus" ($\sqrt{2}$ times the standard deviation). A difference of twice the modulus was considered significant, and differences of 1.5 times the modulus were noteworthy. Edgeworth's historical importance regarding the current topic appears not to lie in his development of a test of statistical significance, but rather in his position as a sort of intermediary between Francis Galton and Karl Pearson. Pearson was clearly influenced by Galton's ideas, but he apparently came to fully appreciate those ideas only after his association with Edgeworth in the early 1890s.

It became evident that Pearson was rather highly motivated by his desire to outdo Edgeworth (Stigler, 1986, p. 338). Pearson gave a series of lectures in 1893 that emphasized Edgeworth's significance testing methods, but, perhaps as a result of their competitive relationship, Pearson decided to measure differences not in terms of the modulus, but rather in terms of a new measure of variation which he called the standard deviation.

Pearson continued to correspond with Edgeworth over the next several years as he worked on issues such as the randomness of the roulette wheel and skew curves. His research and theorizing led to the development of the chi-square goodness of fit test and the birth of modern statistical significance testing. This test was the first to allow for determination of the probability of occurrence of discrepancies between observed and expected frequencies (Cowles & Davis, 1982, p. 555).

Rejection levels had previously appeared (*e.g.*, six times the probable error and 1.5 times the modulus), but with the advent of the chi-square test statistic, levels of rejection began to be standardized. Pearson himself saw the .1 level as "not very improbable" and the .01 level as "very improbable" (Pearson, 1900). William Gosset, who, under the pen name of "Student" developed the t distribution for small samples, determined that a level of three times the probable error would usually be considered important (Student, 1908, p. 13). Wood and Stratton (1910, p. 433) advised agricultural researchers to take "30 to 1 as the lowest odds which can be accepted as giving practical certainty that a difference is significant." It might be noted that "practical certainty" was interpreted as "enough certainty for a practical farmer." Thirty to one odds translate to p=.0323 or a mean difference of 3.2 probable errors. McGaughy (1924) stated that 3 times the probable error (3PE) "is the accepted standard for the undoubted significance of an obtained difference between averages." 3PE is equal to two standard deviations which in turn equals about 4.56%. Cowles and Davis (1982, p. 557) hypothesize that Fisher rounded this figure to .05 to express the significance level in the metric of the standard deviation rather than that of the probable error.

Thus, it is clear that although R. A. Fisher is often credited with the establishment of the .05 level of significance, given his statements, "It is convenient

to take this point as a limit in judging whether a deviation is to be considered significant or not" (Fisher, 1925, p. 47), and "It is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard" (Fisher, 1935/1951, p. 13), this convention was clearly built on the foundations laid by earlier researchers. Of course, there were exceptions to the use of the .05 level. J. E. Coover, an early parapsychological researcher, was, in 1917, unwilling to accept a p value of .00476 as a "decisive indication of some cause beyond chance" (Coover, 1975, p. 82). Despite the exceptions, the .05 level of significance had numerous historical precedents and was anything but arbitrary.

It is important to note, however, that none of these early significance testers asserted that a given level was to be used in all cases. Each researcher at least implied the subjective nature of the choice of rejection level, and Fisher himself encouraged scientists to consider situation-specific circumstances before determining an appropriate level.

## Twentieth-Century Controversies

Danziger (1990) notes that in the first half of the twentieth century traditional methodology of experimental procedure shifted from an emphasis on single-subject research that focused on experimental control and the *a priori* minimization of error to a focus on treatment group experimentation with comparison of aggregate means and the measurement of error after the fact. Such a shift clearly created an environment in which statistical significance testing flourished. The publication of Fisher's texts (1925, 1935) and the attending widespread availability of small-sample statistical procedures contributed greatly to the popularity of the newer methodologies, and it is no exaggeration to say that Fisher's conception of statistical

significance testing had an enormous influence on the evolution of science in the early part of the century.

Unfortunately, Fisher's writings were often contradictory and elusive. It, therefore, is difficult to ascertain with certainty exactly what Fisher's approach to statistical significance testing was. Kempthorne (1984, p. 303) in fact refers to Fisher I and Fisher II. The basic idea, however, is that significance testing yields a statement about the probability of the null hypothesis in the sample, assuming the sample came from a population in which the null hypothesis is true. As such, a finding of statistical significance is a statement about the degree of confidence one has that the sample came from a population in which the null hypothesis is true.

The logic of statistical significance testing is nothing if not controversial, however, and Fisher's logic soon came to be questioned. Two major schools of thought on the subject developed in the middle of the century. Savage states, "The widest cleft between frequentists is that between R. A. Fisher and those who side closely with him on the one hand and those who more or less associate themselves with the school of Jerzy Neyman and Egon Pearson" (Savage, 1961, p. 577). (The conflict between the frequentists and the Bayesians, while influential on Neyman-Pearson and especially Fisher, is not discussed here.) Essentially, the Neyman-Pearson approach involves a logical choice between rival hypotheses--the null hypothesis and the alternative hypothesis. Accordingly, this approach is often called Hypothesis Testing rather than Significance Testing. Concepts introduced by Neyman and Pearson include Type I and Type II errors, power, and critical values of p. Huberty (1993, p. 318) presented the two methods of testing as follows:

| *Significance Testing (Fisher)* | *Hypothesis Testing (Neyman-Pearson)* |
|---|---|
| 1. State H(0). | 1. State H(0) and H(1). |

| | |
|---|---|
| 2. Specify test statistic (T) and referent distribution. | 2. Specify test statistic (T) and referent distribution. |
| 3. Collect data and calculate value of T. | 3. Specify alpha value and determine rejection region. |
| 4. Determine *P* value. | 4. Collect data and calculate value of T. |
| 5. Reject H(0) if *P* value is small; otherwise retain H(o). | 5. Reject H(0) in favor of H(1) if T value is in the rejection region; otherwise retain H(o). |

The Fisher and Neyman-Pearson philosophies are described in some detail in Oakes, (1986, ch. 6), Salsburg (1990), Seidenfeld (1979, chaps. 2 and 3), and Spielman (1974)--and for a response to Spielman, see Carlson (1976).

Statistics textbooks in use since 1910 reflect the divergence of opinion regarding statistical significance testing. Fisher's approach was dominant in the first half of the century, but the Neyman-Pearson philosophy came to be integrated into the textbook presentations during the years 1935-1950 (Huberty, 1993, p. 323). During what Gigerenzer and Murray (1987) have called "the inference revolution" an almost conspiratorial hybridization of the two philosophies of statistical significance testing was created that, to say the very least, wouldn't have pleased Fisher, Neyman, or Pearson. Application came to be emphasized over theory, and the result was a blending of incompatible concepts. Deprived of a theoretical basis from which to operate, researchers in effect came to make sure that popular concepts (*e.g.*, power, p values, Type I and Type II errors, *etc.*) were all accounted for. Thus, experimental studies in the social sciences all too often were analogous to the Biblical "house built upon the sand." They looked nice but were unable to stand up to the force of theoretical scrutiny.

Gigerenzer (1993) presents an interesting and humorous Freudian analogy in which Neyman-Pearson logic functions as the Superego and works with the

Fishererian Ego to censor the Bayesian Id of the hybrid logic. Gigerenzer writes,

> The metaphor brings the anxiety and guilt, the compulsive and ritualistic behavior, and the dogmatic blindness associated with the hybrid logic into the foreground. Is is as if the raging personal and intellectual conflicts between Fisher and Neyman and Pearson, and between these frequentists and the Bayesians were projected into an "intrapsychic" conflict in the minds of researchers. And the attempts of textbook writers to solve this conflict by denying it have produced remarkable emotional, behavioral, and cognitive distortions (p. 325).

However the issues may be viewed, there is little doubt that statistical significance testing stirred a mighty debate that may be considered to have culminated with Carver's influential article, "The Case Against Statistical Significance Testing" (Carver, 1978) and Meehl's statement that reliance on null hypothesis testing is "one of the worst things that ever happened in the history of psychology" (Meehl, 1978, p. 817). When seen in its historical context, however, (a context, by the way, that is necessary if not sufficient) statistical significance testing was probably (a) inevitable and (b) a useful scientific catalyst despite the fact that it was (c) inevitably misinterpreted. But it is conceivable that the development of statistical significance testing was indeed associated with "one of the worst things that ever happened in the history of psychology"--the abandonment of theoretical foundations in favor of experimental expedience.

Statistical significance testing has not only survived the onslaught of virulent criticism it has faced in the last 20 years, it continues to flourish. The disinterested observer might come to the conclusion, as Oakes (1986, p. 68) posited, that behavioral scientists have been willfully stupid. A number of slightly less incriminatory explanations have been offered. Oakes suggests that inertia and submission to statistical authority, the weakness of proposed alternatives, and the prevailing philosophical climate have allowed statistical significance testing to

continue. Carver (1978) holds that the reasons for its continuance include the common misunderstanding that significance testing is associated with replicability and, secondly, that it is used to determine whether the size of a difference is important. While all of these explanations have merit, the last one is the most compelling. The use of the word "significance" has led to the widespread misconception that statistically significant results are important. While this may be the case on occasion, improbable results are not necessarily important results (Thompson, 1993). Furthermore, indiscriminate reliance on statistical testing procedures in place of individual experimenter value-oriented decisions and more informative techniques is symptomatic of the abandonment of the theoretical foundations upon which science is based and of the hybridization that corrupted statistical significance testing in the first place. Calculation of a test statistic that allows a researcher to accept or reject a null or alternative hypothesis does not relieve that researcher of the obligation to differentiate between probability and importance. That obligation has all too often been abdicated.

## Contemporary Practice

Carver (1993) gives four suggestions concerning the current practice of statistical significance testing. They are as follows:

1. Insist that "statistically" be inserted in front of "significant" in research reports.
2. Insist that the results always be interpreted with respect to the data first, and statistical significance, second.
3. Insist that attention be paid to the size of the effect, whether it is statistically significant or not.
4. Insist that new journal editors present their views on statistical significance testing prior to their selection.

Thompson presents a simple model of interpreting experimental results that includes statistical significance testing. The researcher addresses two basic questions-

-"Do I have anything?" and "If so, where does it come from?" Statistical significance testing may be used as an indicator, albeit a relatively unimportant one, for the first question. More important indicators are effect size (via $r^2$ analogs and/or standardized differences) and replicability. As Carver (1978) noted, statistical significance testing is often associated with replicability, but such an association is ill-conceived. Replicability is more appropriately addressed through actual replication or through procedures such as cross-validation as well as bootstrap and jack-knife techniques. Indicators for Thompson's second question include beta weights, structure coefficients, etc.

## Summary and Conclusion

This paper has examined the historical origins of statistical significance testing. Precursors that set the stage for the development of statistical significance testing were explored, and the advent of modern procedures was traced. Current controversies regarding statistical significance testing were discussed, and suggestions for contemporary practice were given.

Serlin (1987, p. 367) stated, "modern reconstructions of science indicate that progress in science is never that self-evident, in that the detection of progress must take on a historical perspective." At the beginning of this paper it was suggested that examining statistical significance testing within its historical perspective might clarify whether current practices evolved from deliberative judgment or merely developed as happenstance that became reified in routine. It is somewhat of a relief to find that the development of statistical significance testing has involved a degree of deliberative judgment. Perhaps it is an idea whose time came and is now gone. But there can be no doubt that statistical significance testing served as an important catalyst for the growth of science throughout the twentieth century.

## References

Arbuthnot, J. (1710). An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. Philosophical Transactions of the Royal Society of London, 27, 186-190.

Carlson, R. (1976). Discussion: The logic of tests of significance. Philosophy of Science, 43, 116-128.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61 , 287-292.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Coover, J. E. (1975). Experiments in psychical research. New York: Arno Press. (Originally published 1917).

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. American Psychologist, 37, 553-558.

Danziger, K. (1990). Constructing the subject. Cambridge: Cambridge University Press.

Ebbinghaus, H. (1885). Uber das Gedächtnis. Translated, 1913 as Memory: A Contribution to Experimental Psychology, trans. Henry A. Ruger and Clara E. Bussenius. New York: Teachers College, Columbia University. Reissued, 1964; New York: Dover Press.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935). The design of experiments. Edinburgh: Oliver & Boyd.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In

G. Keren & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, G., & Murray, D. J. (1987). Cognition as intuitive statistics. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, G., Swijtink, Z., Porter, T., Datson, L., Beatty, J., & Krüger, L. (1989). The empire of chance: How probability changed science and everyday life. Cambridge: Cambridge University Press.

Huberty, C. J. (1987). On statistical testing. Educational Researcher, 16, 4-9.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. Journal of Experimental Education, 61, 317-333.

Kempthorne, O. (1984). Statistical methods in science. In P. Rao & J. Sedransk (Eds.), W. G. Cochran's impact on statistics. New York: John Wiley & Sons.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Laplace, P. S. (1823). De l'action de la lune sur l'atmosphere. Annales de Chimie et de Physique, 24, 280-294.

McGaughy, J. R. (1924). The fiscal administration of city school systems. New York: Macmillan.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Michell, J. (1767). An inquiry into the probable parallax and magnitude of the fixed stars from the quantity of light which they afford us, and the particular circumstances of their situation. Philosophical Transactions of the Royal Society, 57, 234-264.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). The significance test controversy. Chicago: Aldine.

Oakes, M. (1986). Statistical inference. New York: Wiley.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50, 157-175.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. American Psychologist, 46, 1086-1087.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Salsburg, D. (1990). Hypothesis versus significance testing for controlled clinical trials: A dialogue. Statistics in Medicine, 9, 210-211.

Savage, L. J. (1961). The foundations of statistics reconsidered. In J. Neyman (Ed.), Proceedings of the fourth symposium on mathematical statistics and probability (Vol. 1, pp. 575-586). Berkeley, CA: University of California.

Seidenfeld, T. (1979). Philosophical problems of statistical inference. Boston: Reidel.

Serlin, R. C. (1987). Hypothesis testing, theory building, and the philosophy of science. Journal of Counseling Psychology, 34, 365-371.

Spielman, S. (1974). The logic of tests of significance. Philosophy of Science, 41, 211-226.

Stigler, S. M. (1986). The history of statistics: The measurement of uncertainty before 1900. Cambridge, Ma: Belknap Press.

Student (W. S. Gosset). (1908). The probable error of a mean. Biometrika, 6, 1-25.

Thompson, B. (1993). Foreword. Journal of Experimental Education, 61, 285-286.

Thompson, B. (1993). The case against (only doing) statistical significance testing. (An ERIC/Clearinghouse on Assessment and Evaluation Digest). Washington, DC: ERIC.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Walker, H. M. (1929). Studies in the history of statistical method. Baltimore: Williams and Wilkins. Reprinted, 1975; New York: Arno Press.

Wood, T. B., & Stratton, F. J. M. (1910). The interpretation of experimental results. Journal of Agricultural Science, 3, 417-440.